Automating Data Entry for Biomedical Databases

G.R.Thoma
National Library of Medicine
Bethesda, Maryland 20894, USA

SUMMARY

Data entry from paper-based literature, e.g., medical journals, into databases has traditionally been manual and labor-intensive. As the journal sets grow in number and the number of entries correspondingly increase, the traditional approach has proven to be burdensome. More automated approaches relying on scanning, optical character recognition (OCR), image segmentation, biomedical lexicons and image processing have to be introduced.

In response to a crisis in the data entry of citations and abstracts of medical journal articles for the MEDLINE database, the R&D center at the National Library of Medicine developed a system, temporarily code named MARS for Medical Article Record System, that combines keyboarding of citation data (journal name, date, author, title, affiliation, page numbers, etc.) with scanning and OCR text converting abstracts. The abstract is usually the largest part of a citation record, and keyboarding it is generally labor-intensive.

The initial MARS, currently in operation, consists of multiple workstations of three types: scanner, proofing and editing, and keyboarded citation entry. In addition, the system requires three servers: a network file server, an OCR server and one to match double keyboarded citations. To select an appropriate OCR system, the performance of six OCR packages were compared in terms of detected (highlighted) errors, highlighted correct words (false alarms) and undetected errors. The selection of the OCR package for the MARS system was based on minimizing the undetected error rate since this is the factor that gives a high level of confidence that the proofers need not compare the converted text against the original printed material, a highly labor-intensive step.

Briefly, the MARS system works as follows. The scanning operator scans the pages on which the abstracts appear, and zones the titles and abstracts, and the bitmapped TIFF files are sent to the network server. The OCR server performs text conversion, and produces a text file of the abstract and title. Concurrently or at any time, a citation entry operator keys in all the fields in a template for the journal issue and each article, and a second operator repeats this process for the same journal issue and articles. A CITATION MATCH module in the network server compares the two citation entries, and produces a "citation difference" file highlighting inconsistencies. A MATCH module in the server then matches this difference file and the OCR'ed abstract, correlating the two article title texts (one from the scanned page and the other from the keyboarded citation). At this point, the combined file is available for validation and proofing. Following this step, the file is FTP'ed to a module in the NLM mainframe computer. The completed record files in the mainframe are later accessed by indexers who add the appropriate descriptive information such as Medical Subject Headings.

1063-7125/98 $10.00 © 1998

All workstations are networked via a LAN. One advantage of this networked approach is that all three functions (scanning, proofing/editing and citation entry) can be done concurrently, so that a fixed sequence of operations is not necessary. For example, citations may be keyboarded before or after abstracts are converted. The network server maintains directories in which the scanned TIFF images, the text abstract files and the citation files are all kept till they are acted upon.

A year after the system was placed in operation, it was providing over 20% of the total entry requirements of NLM. Our current work is to design a database-centered system that will provide more comprehensive automation, and lower per-unit cost, by incorporating subsystems for autozoning, page segmentation, automatic field identification, and automated syntax reformatting.